VIABLE NASH EQUILIBRIA: FORMATION AND DEFECTION
(FEB 2020)

EHUD KALAI

In memory of John Nash

ABSTRACT. To be credible, economic analysis should restrict itself to the use of only those Nash equilibria that are viable. To assess the viability of an equilibrium π, I study simple dual indices: a formation index, $F(\pi)$, that specifies the number of loyalists needed to form π; and a defection index, $D(\pi)$, that specifies the number of defectors that π can sustain.

Surprisingly, these simple indices (1) predict the performance of Nash equilibria in social systems and lab experiments, and (2) uncover new properties of Nash equilibria and stability issues that have so far eluded game theory refinements.

JEL Classification Codes: C0, C7, D5, D9.

1. OVERVIEW

Current economic analysis often relies on the notion of a Nash equilibrium. Yet there are mixed opinions about the viability of this notion. On the one hand, many equilibria, referred to as *viable* in this paper, play critical roles in functioning social systems and perform well in lab and field experiments.

Date: March 9, 2018, this version Feb 24, 2020.
Key words and phrases. Normal form games, Nash equilibrium, Stability, Fault tolerance, Behavioral Economics.
The author thanks the following people for helpful conversations: Nemanja Antic, Sunil Chopra, Vince Crawford, Kfir Eliaz, Drew Fudenberg, Ronen Gradwohl, Yingni Guo, Adam Kalai, Fern Kalai, Martin Lariviere, Eric Maskin, Rosemarie Nagel, Andy Postlewaite, Larry Samuelson, David Schmeidler, James Schummer, Eran Shmaya, Joel Sobel, and Peyton Young; and seminar participants at the universities of the Basque Country, Oxford, Tel Aviv, Yale, Stanford, Berkeley, Stony Brook, Bar Ilan, the Technion, and the Hebrew University. This paper replaces earlier versions of similar titles.

Unviable Nash equilibria, on the other hand, fail to perform well in such experiments and are not observed in functioning social systems.[1] As a result, the uses of unviable Nash equilibria in many recent theory and applied theory papers seem contrived and unproductive.

This paper focuses on the subject of *equilibrium viability*. As a starting point, we study simple dual economic indices that succeed in assessing the viability of an equilibrium π of an n- person strategic game Γ: a *formation index*, $F(\pi)$, that reflects the difficulty of forming the equilibrium π; and a *defection index*, $D(\pi)$, that reflects the ability of a formed equilibrium π to sustain defection. As illustrated by examples in the paper, viable Nash equilibria perform well on these indices, whereas unviable equilibria perform poorly. The indices also identify new issues concerning Nash equilibrium stability that so far have eluded game theory refinements.[2]

Our view is that viability assessments are subjective, and it is better to base them on multivalued indices than on binary game-theory refinements. This view is illustrated through the use of $E(X)$ and $SD(X)$ by an economic agent who wishes to assess the viability of an investment X. While no index or refinement tells her whether or not the investment X is viable, $E(X)$ and $SD(X)$ are useful for forming a subjective assessment of X's viability. The subjective part of the assessments enables one to consider additional information such as the history, context, and objective of an equilibrium; and other considerations that are not part of the formal description of the game.

[1]The literature on these issues is too large to survey here. Some key examples and references may be found in Smith (1982), Erev and Roth (1998), Crawford (1998), Kahneman and Tversky (2000), Goeree and Holt (2001), Camerer (2003), and many of their follow-up papers.

[2]Some key examples are Aumann (1959), Selten (1975), Myerson (1978), Basu and Weibull (1991), Kreps and Wilson (1982), Kalai and Samet (1984), Kohlberg and Mertens (1986), Bernheim, Peleg, and Whinston (1987), Young (1993), Kandori, Mailath, and Rob (1993), Moreno and Wooders (1996), and Myerson and Weibull (2015).

This paper deliberately makes a minimal departure from the concept of Nash equilibria. It keeps the anonymity and ordinality properties of Nash equilibria.[3] Its only deviation from the concept of Nash equilibria is in eliminating the assumption that "no opponent is a potential defector" and replacing it by a single variable that describes the "number of opponents who are potential defectors." This single departure means that all the new observations made in the paper are due solely to the inclusion of this particular variable. Moreover, the minimality of the departure means that the indices stay away from various refinements of Nash equilibria and thus (1) remain relevant to a broad range of applications, and (2) rely on simple best-response computations that are comprehensible to most players.

In the concluding section we point to applications that call for more refined viability indices. However, since such refinements may require different and inconsistent modifications to the notion of Nash equilibrium, their study is left for future research.

1.1. **RELATED LITERATURE.** Low viability may be viewed as a non-binary refinement of Nash equilibrium. As such, it presents concerns about stability that existing game theory refinements have not addressed. More specifically, this paper presents examples of equilibria that pass most refinement tests, but would be considered to have low viability in the context in which they are played.

Due to the natural interpretations of the indices in strategic games, it is not surprising that they appear in earlier game theoretic discussions.

Formally, the D index may be viewed as the *level of subgame perfection*[4] in a two-stage play of π, in which the players are allowed to revise their chosen

[3]Nash asks only whether some player has an incentive to defect; he has no concerns about the defector's identity and her role in the game, or about the (cardinal) strength of her incentives to defect.

[4]See Kalai and Neme (1992)

strategies after learning the choices of their opponents in stage one. Choosing π in the first stage and playing the second stage with no revisions is k subgame-perfect if and only if $k < D(\pi)$.[5]

In applications, D has its roots in the computer science literature on distributive computing; see, for example, Ben-Or et al. (1988). Motivated by this literature, Eliaz (2002) provides an implementation procedure for economic environments with *faulty players*. To do so, Eliaz introduces the concept of a *k-fault-tolerant equilibrium*; in section 9 below, we show that an equilibrium π is k-fault-tolerant if and only if $k < D(\pi)$.

In addition to Eliaz (2002), two other papers pertinent here are Abraham et al. (2006) and Gradwohl and Reingold (2014). The level of *resilience* used by Abraham et al. (2006) in their implementation of secret-sharing and distributive-computing games is the same as the index D used in this paper. And Gradwohl and Reingold (2014) develop a notion of fault tolerance that holds in equilibria of large games.

The F index is new to this paper, but it is related to the earlier literature through its duality relationship with D, namely, $F = n - D$.

However, this paper deals with applications that cannot be addressed by the papers cited above. An important difference is that in this paper we address the incentives of rational players to defect, in addition to the irrational faulty-players defections discussed in the papers above. As a result, the indices are useful in addressing questions such as when rational players choose to defect, how to modify a game to turn potential defectors into loyalists and vice versa, and how to form or switch an equilibrium in a game played by rational players.

[5]More explicitly, in stage one every player i declares the strategy (pure or mixed) ς_i that she *intends* to play. In stage two, with full knowledge of the entire profile of intentions ς, she chooses her *actual* strategy $\sigma_i = \sigma_i(\varsigma)$. The payoff of the two-stage game is $u_i(\varsigma, \sigma) = u_i(\sigma)$, if her $\varsigma_i = \pi_i$; otherwise $u_i(\varsigma, \sigma) = w$, where w is the lowest possible payoff of any player in Γ. Consider the strategy profile $\overline{\pi} = (\pi, \pi^e)$ in which $\varsigma = \pi$, and $\pi_i^e(\varsigma) = \varsigma_i$. It follows directly from the definitions that π is k-perfect iff $k < D(\pi)$.

2. Viability Indices

2.1. **GAME THEORETIC PERSPECTIVE.** A game theory story that motivates the definitions of the indices below concerns n rational players about to play a focal strategy profile π. However, each of these rational players is concerned that a number of her opponents may fail to play their π strategies (despite their seeming optimality) and asks herself whether, in view of her concerns, her π strategy is optimal. This is a difficult question since there are many such concerns, and they are often difficult to formulate. One concern is that some of her opponents are simply "faulty," i.e., irrational and unpredictable as described by Eliaz (2002). But she may even have concerns about the strategies chosen by *rational opponents*. Examples include profitable defections by groups of rational opponents, outside bribes and threats applied to rational opponents, rational opponents' desire to build their reputations for future games, as well as many other concerns not included in the formal description of the game under consideration.

Addressing the concerns above by standard Bayesian methods is impractical. It would require listing the large number of such concerns that each player holds about each of her opponents, and assigning a prior probability to each such concern.

Notice, however, that these concerns become irrelevant when the player's π strategy is dominant in the game under consideration, since her strategy is the best regardless of the strategies her opponents end up choosing. For this reason, a profile π that consists of all dominant strategies (i.e., a dominant-strategy equilibrium) is most viable according to the indices of this paper, i.e., $F(\pi) = 0$ and $D(\pi) = n$. But when π is not a dominant-strategy equilibrium, we may ask a related and less demanding question: against how many potential defectors are the players' strategies dominant? Such questions underlie the definitions of $D(\pi)$ and $F(\pi)$ below.

2.2. **GAME-THEORETIC NOTIONS.** This paper focuses on a strategic n-person game $\Gamma = (N, A = \times_{i\in N} A_i, u = (u_i)_{i\in N})$. N is a finite set of n *players*. Subsets of N are called *groups* or *coalitions,* and N is the *grand coalition*. Elements of A_i are the *strategies of player* i, and strategy *profiles*, or *profiles* for short, are functions of the form $\alpha = (\alpha_i)_{i\in N} \in A \equiv \times_i A_i$, which assign to every player i an element $\alpha_i \in A_i$. We assume that Γ has no *dummy players*, i.e., $|A_i| \geq 2$.[6]

For a group G and a profile α, define $\alpha_{_G} = (\alpha_j)_{j\in G}$. The profile in which the players in G defect from α to a profile β is defined by: $(\alpha_{N\setminus G} : \beta_G)_j = \alpha_j$ for $j \notin G$, and $(\alpha_{N\setminus G} : \beta_G)_j = \beta_j$ for $j \in G$. A strategy α_i of player i is a best response to a profile β, if $u_i(\beta_{N\setminus i} : \alpha_i) \geq u_i(\beta_{N\setminus i} : \chi_i)$ for any strategy χ_i of player i.[7].[8] A profile α is a best response to a profile β if every α_i is a best response to β; and a profile α is a Nash equilibrium if it is a best response to itself.

A strategy α_i is dominant if α_i is a best response to any profile; and a profile α is a dominant-strategy equilibrium if each α_i is dominant. The notions of best response and domination are weak, in the sense that they are defined by weak inequalities.

Throughout the rest of the paper, $\pi = (\pi_i)_{i\in N}$ denotes one *arbitrary, fixed focal profile*. The implicit assumption that π is focal explains why we may think of π as an equilibrium in cases in which best-response functions involve indifference. Similarly, the definitions of the indices below are motivated by this implicit assumption.

Given the fixed profile π, for any profile β and player i, i is a π-defector at β, or a *defector* for short, if $\beta_i \neq \pi_i$; and i is a π-loyalist at β, or a *loyalist* for short, if $\beta_i = \pi_i$.

[6]|S| denotes the number of elements in a set S.
[7]When it is clear from the context, we sometimes omit the brackets and replace $\{i\}$ by i .
[8]For the obvious reasons, whether α_i is a best response to β is independent of the β_i coordinate.

For a profile χ and a group G, it is useful to consider the (sub)game induced on the members of G when the players outside G are committed to their χ strategies.

Definition 1. ***The game played by*** G ***under*** χ, Γ_G^χ , *is defined as follows: the set of players is* G*; the strategy set of every* $i \in G$, A_i, *is the same as in* Γ*; and the payoff of every player* $i \in G$ *at any profile* α_G *is the same as her payoff in* Γ *at the concatenated profile* $(\chi_{N\setminus G} : \alpha_G)$.

Below is an example of an asymmetric game, used to illustrate the concepts defined in the following sections.

Example 1. ***The Party Line Game:*** *Simultaneously, each of three Democrats and five Republicans selects one of two choices,* E *or* F*. The payoff of a player is the number of opposite-party players whose choice she mismatches. We will consider the* divisive equilibrium*, Div, in which all the Democrats choose* F *and all the Republicans choose* E*.*

2.3. **DEFINITIONS AND PROPERTIES OF THE INDICES.** Even though several of the definitions and properties below are dual and equivalent, they highlight different useful aspects of the indices. Some examples in the paper illustrate uses of $D(\pi)$ in addressing coalition defections, $F(\pi)$ in addressing equilibrium formation, and the Nash critical mass $NCM(\pi)$ in Eliaz (2002) implementation. It is natural to think of each of the indices as the primitive in its own set of applications.

The reader should be aware that the concepts and the relationships described below involve significant subtleties, and that the short combinatorial proofs of the propositions are the consequence of careful definitions and the order of presentation chosen. For example, the final argument that $NCM(\pi) = F(\pi) + 1$ makes careful use of the original definition of $D(\pi)$, the dual relation of $D(\pi)$ with $F(\pi)$, and the interpretation of $F(\pi)$ as the formation index. In addition,

this argument and other arguments throughout this section make repeated use of the anonymity and coalitional-monotonicity properties in the notions of defection deterrence and equilibrium formation. In turn, this monotonicity is a result of the anonymity and domination properties used in the definitions of defection deterrence and formation.

2.3.1. ***The defection index: confidence in individual strategies.*** The next definition explains why a high value of $D(\pi)$ assures every player i that her π_i strategy is optimal, no matter what strategies a large number of potential defectors end up playing.

Definition 2. *The defection index $D(\pi)$ is the smallest integer $d = 0, 1, 2, ..., n$ such that π is not a best response to some d-defectors profile α . If π is a best response to all profiles α, then define $D(\pi) = n$.*

For example, the divisive equilibrium of the previous section has $D(Div) = 2$. To see this, observe that if two Democrats choose E instead of F, then E is no longer a best response of any Republican, so $D(Div) \leq 2$; and since no single player's choice can motivate any player to defect from Div, it follows that $D(Div) > 1$.

2.3.2. ***Deterring multiplayer defections.*** Under the equivalent definition below, $D(\pi)$ expands the Nash equilibrium property that any single defector (weakly) loses, to the property that every member of any group G of defectors loses, for all Gs with $|G| \leq D(\pi)$.

Definition 3. Defection deterrence*:*

1. The profile π (strongly) deters defection of a group G *if π_g is a dominant strategy in the (defection) game Γ_G^π for every player $g \in G$.*

2. The profile π (strongly) deters d defectors*, if it deters the defection of any d-player group G.*

Item 2 above requires that π deters the defection of the players in any known group of d potential defectors G. It is straightforward to formulate a Bayesian game with d-potential defectors, B_d, that shows that such a player is deterred even if she has no information about the identity of her potential codefectors.[9]

The word *strongly* was inserted (in parentheses) in the definition above to emphasize that π_g is required to be a dominant strategy and not just an optimal one. This distinction does not come up in the case of single defectors treated by Nash, but it must be addressed when we discuss multiplayer defections. The choice of strong deterrence means that a high value of D is a strong condition and that a low value may require further examination.

Notice also that the anonymity and domination properties in the definition above imply a monotonicity in group deterrence: if π deters d defectors, then it deters d' defectors for any $d' < d$.

The next proposition shows that $D(\pi)$ can be equivalently defined to be the maximal number of defectors that π deters.

Proposition 1. Defection Deterrence*: The profile π deters d defectors if and only if $d \leq D(\pi)$.*

Proof. Let G be any group of size $|G| < D(\pi)$ and consider any player $g \in G$. If π_g is not dominant in Γ^{π}_{G}, then there is a Γ profile α, with fewer than $D(\pi)$ defectors, to which π_g is not a best response. This contradicts the minimality condition in the definition of $D(\pi)$. So π deters the defection of any group G with $|G| < D(\pi)$.

[9] B_d starts with a random draw of a group G of d potential defectors according to a prior probability distribution μ. Every member of G is informed only of her selection and is asked to select a defecting strategy from her set of strateagies A_i. Her final payoff is the payoff computed at the underlying game Γ, when all the selected defectors use their defecting strategies and all the other players play their π strategies. Notice that in B_d each player has only one information set; thus, its play consists of a single choice of a strategy $\alpha_i \in A_i$ for every player i. It is straightforward to verify that π deters d defectors as defined above if and only if π is a dominant (defecting) strategy equilibrium in the Bayesian defection game B_d for every prior μ.

For the converse, observe that if π deters d defectors, then for every player i, π_i is a best response to any profile α with $d-1$ or fewer defectors. Since this is true for every player i, π is a best response to any profile α with $d-1$ or fewer defectors, i.e. $D(\pi) > d-1$, or $D(\pi) \geq d$. □

In the *Div* equilibrium, for example, no member of a pair of potential defectors can imagine gaining from her defection, regardless of the strategies she ascribes to her fellow defector; so *Div* deters two defectors. But *Div* may fail to deter three defectors. For example, if one Republican and two Democrats are potential defectors, then the single Republican can imagine that, if the two Democrats choose E he is better off switching to F.

2.3.3. ***The formation index and duality.***

Definition 4. The formation index *of* π *is defined by* $F(\pi) = n - D(\pi)$.

Next we show that the formation index can be equivalently defined to be the smallest number of players that can form π, and that it is dual to the defection index.

Definition 5. Equilibrium formation*:*

(1) *A* group L forms π *if* π_i *is a dominant strategy in the game* $\Gamma^{\pi}_{N\setminus L}$ *for every* $i \in N\setminus L$.
(2) *For any* l, l *(players)* form π, *if any* l*-player group forms* π.

Lemma 1. Duality*: The profile* π *deters* d *defectors if and only if* $n-d$ *players form* π.

Proof. As is clear from the definitions of the two concepts, any group L forms π iff π deters the defection of the group $N\setminus L$. This implies that any group of l-player group forms π iff π deters defection of any group of $(n-l)$ players. □

As is the case for group defection deterrence, the anonymity and domination properties in the definition of group formation imply a monotonicity in group formation: l players form π implies that l' players form π for any $l' \geq l$.

Proposition 2. Equilibrium Formation*: l players form π iff $l \geq F(\pi)$.*

Proof. $l \geq F(\pi)$ iff $n-l \leq n-F(\pi)$, i.e., iff $n-l \leq D(\pi)$. By the last Lemma, the last condition is equivalent to l players forming π. □

At the divisive equilibrium of the Party Line game, for example, any six players who commit to play Div make it a dominant choice for each of the remaining two players to play Div. This is so because any group of six players must either (i) include all the players of one of the two parties, or (ii) have a majority in both parties; in either case their commitment to play their Div strategies makes the Div strategies dominant for each of the two remaining players. But a five-player commitment to play Div may not be sufficient for Div formation. For example, the commitments of one Democrat and four Republicans to their Div strategies does not make E a dominant strategy for the fifth Republican.

As should be clear, our definitions are restricted to anonymous one-shot equilibrium formation. The reader is referred to our discussion of future research for an elaboration on the complex subject of nonanonymous dynamic equilibrium formation.

2.3.4. ***Nash critical mass, small worlds, and faulty players.*** *A small-world property* of Nash equilibria (see Mertens (1992)) provides an alternative characterization of the formation index. This property states that an equilibrium π defined for a "large-world game" that consists of n players can be played in isolation in a "smaller-world game" that consists of a subset of the players. More specifically, Nash incentives in the small world hold, no matter what the players outside the small world choose to play. The concept of *Nash*

fault tolerance, used by Eliaz (2002), is an equivalent condition that illustrates a strong economic application of this condition. To describe the condition we use a concept of a uniform Nash equilibrium.

Definition 6. *For any group G, π is a* uniform Nash equilibrium *of G, if π is a Nash equilibrium of the game Γ_G^χ for every profile χ.*

The Nash critical mass *of π, $NCM(\pi)$, is the smallest integer b such that π is a uniform Nash equilibrium of any group G with $|G| \geq b$.*

Proposition 3. Nash critical mass of a small world*:* $NCM(\pi) = F(\pi) + 1$.

Proof. If $|G| \geq F(\pi)+1$, then for every $g \in G$, it is the case that $G\backslash g$ forms π. Therefore π_g is g's best response to π in the game Γ_G^χ for any χ. This shows that π is a uniform Nash equilibrium of G for any G with $|G| \geq F(\pi) + 1$.

For the converse, observe that there is a $D(\pi)$-defectors profile α with a player i for whom π_i is not a best response. Consider the two cases: (1) $\alpha_i = \pi_i$, and (2) $\alpha_i \neq \pi_i$. Without loss of generality, we can change α_i to be π_i to obtain the existence of a profile α with a number of loyalists $\geq F(\pi)$, to which π_i is not a best response for one of the loyalists. This means that we can find an $F(\pi)$-loyalists group G with a member i for whom π_i is not a best response to α. In other words, π is not a uniform Nash equilibrium of G. Since G has $F(\pi)$ members, $NCM(\pi)$ must be greater than $F(\pi)$. □

For example, the Nash critical mass of the divisive equilibrium of the Party Line game is $NCM(Div) = 7(= F(Div) + 1)$. It is a Nash equilibrium for any seven players to "follow the party line," even if they are not sure who the excluded player is and what she may play. But in a six-player game played by all the Republicans and one Democrat, following the party line may fail to be a uniform Nash equilibrium. For example, the Republicans are not best responding if the two excluded players are Democrats who choose E.

It is important to emphasize that the statement "l players play π" is qualitatively stronger for $l \geq F(\pi) + 1$ than for $l \geq F(\pi)$. The latter statement means that the l players incentivize the remaining $n - l$ players to play their π strategies, but it does not address the incentive of the l players to do so themselves. On the other hand, $l \geq F(\pi) + 1$ means that in addition to incentivizing the others, the l players themselves have Nash-equilibrium incentives to play their π strategies.

2.3.5. ***Nash equilibrium progressions.*** The properties and observations discussed above introduce a view of Nash equilibria of n-person games as a progression formed by the rungs in a ladder of sustainability/domination. The least sustainable are profiles π that are not Nash equilibria with $D(\pi) = 0$, i.e., they fail to deter some single-player defectors. Progressively, the next levels correspond to Nash equilibria π with $D(\pi) = 1, 2, ..., n - 1$, where $D(\pi) = d$ includes the equilibria that deter up to d defectors. The most sustainable are the dominant-strategy equilibria with $D(\pi) = n$, which deter any number of defectors.

Notice that this is also a classification that presents a decreasing progression in domination. At the top are the dominant-strategy equilibria π, with $F(\pi) = 0$. Their formation requires the commitment of no players. Progressively, the next levels correspond to Nash equilibria with $F(\pi) = 1, 2, ..., n - 1$, where $F(\pi) = f$ includes the equiliria that require the commitment of at least f players to make π a dominant strategy for the rest. $F(\pi) = n$ includes the remaining non-Nash-equilibrium profiles, in which the commitment of all n players is needed to secure the play of π.

The Ride Sharing game, Example 7, shows that all the levels of these progressions, $0, 1, ..., n$, are obtained in simple games that require only reasonable levels of computational ability.

Proposition 4. Classification of Nash equilibria*: π is not a Nash equilibrium iff $D(\pi) = 0$ (or $F(\pi) = n$); π is a dominant-strategy equilibrium iff $D(\pi) = n$ (or $F(\pi) = 0$); and the intermediary values $D(\pi) = 1, ..., n-1$ (or $F(\pi) = n-1, ..., 1$) partition all remaining Nash equilibria into increasing levels of defection deterrence (or increasing ease of formation).*

Proof. From the definition of D, π is not a Nash equilibrium iff $D(\pi) = 0$, and thus π is a Nash equilibrium iff $D(\pi) > 0$. Also from the definition of D, π is a dominant-strategy equilibrium iff $D(\pi) = n$. This means that the remaining intermediary values must be assigned to all non-dominant-strategy Nash equilibria. □

3. Subjective viability assessments

The first two examples contrast a high D-value equilibrium with one of low D-value.

Example 2. *A **(Language) Matching Game**.*

Simultaneously, each of $200M$ players selects one option (say, a language) from a set of possible choices. For any choice X, the payoff of a player who chooses X equals the number of opponents she matches, i.e., the number of other players who also choose X.

Consider the profile eE in which every player chooses E. It is easy to see that eE is a Nash equilibrium; but beyond the Nash condition, eE deters defections of groups of players. For example, for any player who is one of $1M$ potential defectors, staying with E is a dominant strategy: she would match at least $199M$ opponents by choosing E and match at most $1M - 1$ opponents if she defects. This group deterrence property holds for any group of d potential defectors for ds up to $100M$.

In contrast to the large defection-deterrence value $D = 100M$ above, the defection deterrence of the next equilibrium is only $D = 1$, barely enough to be classified as a Nash equilibrium.

Example 3. *A **Confession Game** (a stag-hunt game that only sounds like a prisoner's dilemma).*[10]

Simultaneously in separate rooms, 36 partners in a crime are interrogated by the police. If none of the suspects confesses, everyone will be released with no penalty. However, if one or more confess, then every suspect will be sentenced to ten years in prison, except for the confessors, who will be sentenced to only three years instead of ten.

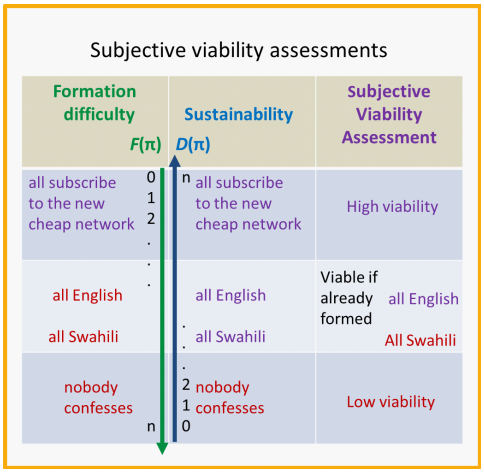
The cooperative profile nC in which nobody confesses is a Nash equilibrium. However, nC fails to deter (with certainty) the defection of a player who believes that at least one more player is a potential defector: if another player confesses, than she is better off confessing herself. Since nC deters sole potential defectors, but fails to deter defection in the presence of one more potential defector, it is a Nash equilibrium but with the low D value , $D(nC) = 1$.

Next, we focus on newly proposed equilibria and contrast one with a high F-value with one with a low F-value.

For the high F-value, consider the language-choice game above, but with an equilibrium in which everybody chooses Swahili, eS. Just like the eE equilibrium in the game, $D(eS) = 100M$, and by the duality of the two indices, $F(eS) = 200M - D(eS) = 100M$. So it would take $100M$ players to choose S, to make S a dominant strategy for the others.

In contrast to the difficult-to-form equilibrium just above, with $F(eS) = 100M$, the next game illustrates an equilibrium that is easy to form; its F value is 10.

[10]The author thanks Adam Kalai for suggesting this example.



Example 4. *A **New Communication Network** is offered for subscription in a population of* $200M$ *potential users, at an individual cost of* \$9.99. *The payoff of a subscriber is* $k - 9.99$*, where* k *is the number of opponents who subscribe. The payoff of a nonsubscriber is zero.*

Consider the equilibrium in which everybody subscribes, $eSub$. It is easy and natural to compute the formation index of this equilibrium directly: Subscribing is a dominant strategy if and only if at least ten other players subscribe; thus, $F(eSub) = 10$. Notice that this computation does not even require knowledge of n, which would be needed for the indirect computations $F = n - D$. But now we can use the duality to conclude that if formed, $eSub$ should be highly sustainable, $D(eSub) = 200M - 10$.

The chart titled "Subjective viability assessments" illustrates how one may assess viability for the equilibria discussed above.

The bottom layer of the chart refers to equilibria similar to the nobody-confesses equilibrium, which have a low D value, i.e., even if formed, their sustainability is questionable. In addition, for a reasonable number of players n, they have a relatively high formation index, i.e., they are difficult to form. In our Confession example, $F(nC) = 35$; thus, a player must be highly confident that all her 35 partners in crime will not confess, to adopt the no

confession strategy herself. Given the difficult formation and the low sustainability even if formed, one would reasonably consider nC to be an equilibrium of low viability.

It is also easy to assign viability to equilibria in the top layer of the chart. Our example in which everybody subscribes to the new cheap network, $eSub$, has a formation index $F(eSub) = 10$ and a defection index $D(eSub) = 200M - 10$. An entrepreneur who owns the network should find it easy to form the equilibrium, as he would need to recruit only 10 out of the 200M players to subscribe. Given the relatively easy formation and high sustainability if formed, one would reasonably consider $eSub$ and similar equilibria to be highly viable.

The middle layer in the chart is more interesting. For large values of n we may obtain equilibria that have both a large D value and a large F value. Such equilibria should be difficult to form but also highly sustainable if formed. If such an equilibrium were already formed – for example, due to some historical evolution – it should be highly sustainable. But if it is not yet formed, the difficult formation makes this equilibrium unviable. For example, in the US all-choose-English is a viable equilibrium whereas all-choose-Swahili is not.

4. Viable equilibria in social systems and the lab

4.1. **Highly sustainable equilibria in social systems.** Many conventions and social arrangements in large populations rely on highly sustainable Nash equilibria, similar to the everybody-choosing-English equilibrium above. Some examples of such viable equilibria are: everybody choosing Spanish, Mandarin, or one of many other languages spoken in different populations; everybody using dollars, euros, or other currencies used in various markets; everybody obeying traffic signals; everybody using the same communication software and/or the same hardware; and the choice of market locations made by sellers and

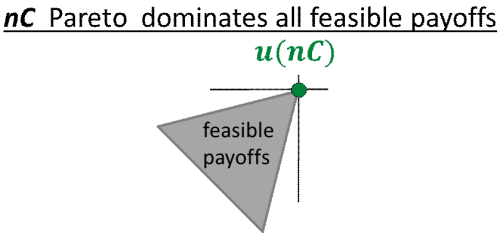
buyers as places to trade. Dress codes, food culture, and many social mores also follow such equilibria.

4.2. **Unsustainable examples in social systems.** With $D(nC) = 1$, players may think of the no-confession equilibrium nC discussed above to be of low viability. Below are three familiar examples of equilibria with $D = 1$ often discussed in the context of social systems and lab and field experiments.

In the **beauty-contest game**, each of n judges submits a real number $r_i \in [0, 100]$. The judges whose submitted number is closest to two-thirds of the average submitted number, $\frac{2}{3}\sum_{i=1}^{n} r_i/n$, are each paid one, while the others are paid zero. The only Nash equilibrium – in which all the judges submit the zero score, $eZero$ – is rarely observed in lab and field experiments. The nonviability of this (strict Nash) equilibrium can be explained by its low defection index, $D(eZero) = 1$. Indeed, the low viability of this equilibrium has given rise to a large literature that offers alternative models of behavior for this and similar games; see Nagel (1995), Crawford et. al. (2013), and Mauersberger and Nagel (2018) for theoretical and empirical references.

Mixed-strategy equilibria are often minimally sustainable, as can be seen again in the Language Matching game. Consider the equilibrium in which every one of the $200M$ players chooses English or Spanish with equal probability. If one player changes her choice probabilities to 2/3 on choosing Spanish and 1/3 on choosing English, then every player's best response is to choose Spanish with probability one. So the mixed-strategy equilibrium has the minimal defection level $D(eE) = 1$. This type of low viability is the case for many mixed-strategy equilibria; we refer the reader to O'Neill (1987) and follow-up papers for empirical studies.

Simple production lines rely on equilibria with a low defection index, $D = 1$. Consider a group of n workers who stand to receive a bonus if and only if they all report to work. A player's payoff is positive if everyone reports



to work and she receives the bonus; it is zero if she does not report to work; and it is negative if she reports to work but receives no bonus because somebody else did not report to work. Having everybody report to work is an appealing equilibrium, but one with a defection index of only 1, since one worker not showing up incentivizes others to also not show up. This low sustainability level is one of the reasons that companies such as Toyota use more sophisticated production lines with higher levels of sustainability (see Mishina (1992)).[11]

5. Comparison with equilibrium refinements

The nobody-confesses profile nC of the Confession Game is a Nash equilibrium with the lowest possible level of defection deterrence, $D(nC) = 1$. However, considerations of standard game theory lead one to conclude that nC is an appealing equilibrium. As one can see in the payoff graph of the two-player case, $u(nC)$ strongly and uniquely Pareto-dominates every feasible payoff in the game. Any defection – whether deterministic or probabilistic, and whether by individuals or groups – is certain to strictly decrease the expected payoffs of all the players, including the defectors. As Aumann and Sorin (1989) argue, this should be the undisputed outcome of this game. Indeed, this equilibrium is *perfect* à la Selten (1975), *proper* à la Myerson (1978), *strong* à la Aumann (1959), and *coalition-proof* à la Bernheim, Peleg, and Whinston (1987).

It seems, however, that crime syndicates are not impressed by nC's high acclaim among game theorists. They are more concerned that somebody

[11]The author thanks Sunil Chopra and Martin Lariviere for providing this reference.

would confess out of fear that others might, a concern that coincides with the reasons for π's low D value. Crime-syndicate remedies, such as killing confessors, change the game's payoff functions to make no-confession a dominant strategy with the maximal D value n. Similar concerns lead more respectable organizations, such as high-tech and bio-research companies, to demand that their employees sign no-disclosure agreements.

The nC equilibrium above also reinforces our earlier point that the reasoning behind our stability index is simple enough to be understood even by players with limited computational ability. Players simply have to fear the outcomes if other players defect.

The next example illustrates a difference between the defection index D and refinements based on evolutionary stability. In particular, an evolutionary stable equilibrium may still be minimally sustainable.

Example 5. ***Match The Center.*** *The boss, B, and n subordinates each selects one element from a given set of choices. B's payoff is 1 if he chooses E, zero otherwise; and every subordinate's payoff is 1 if her choice matches B's choice, zero otherwise.*

The game has a unique equilibrium eE, in which every player chooses E. This equilibrium has a large basin of attraction and it is stochastically stable in the sense of Young (1993) and Kandori et al (1993).[12]

However, eE is only minimally sustainable, i.e., $D(eE) = 1$. All the subordinates are vulnerable because the optimality of their choice depends entirely on the choice of one player, B. Yet they may have concerns about B's commitment to his strategy, his ability to withstand pressure coming from outside the game, his desire to create a reputation for future games, possible miscalculations, and so forth.

[12]For further discussion and references, we refer the reader to Ellison (2000).

It is important to note that, unlike the dynamic approach of evolutionary game theory, our model deals with a game that is played just once. Moreover, in many evolutionary models deviations from equilibrium are controlled by forcing mutation probabilities to approach zero. Here, on the other hand, imagined deviations from equilibrium are controlled only by bounding the number of players who may deviate, but not the probabilities and magnitude of deviations.

As a side remark about centrally controlled games, notice that the sustainability of the eE equilibrium can be increased if the center is occupied by a group of bosses. For example, if there are three bosses, all having E as a dominant strategy, and the subordinates all wish to match the majority choice of the bosses (e.g., a politburo instead of a dictator), then $D(eE) = 2$, and not 1 as it is for a single boss.

6. Information, signals, and signal duplication

Our next example illustrates the fact that duplicating information can be used to increase equilibrium sustainability, even if the duplicated information is common knowledge. The example is a simple version of a Crawford-Sobel (1982) sender-receiver game, in which the senders, their information, and their strategies are simply duplicated.

Example 6. ***Signalling Game with duplicated senders****: A two-element set $T = \{\alpha, \beta\}$ denotes two possible states, two possible signals, and two possible actions in the game. There are three recommenders, of whom two are honest and one is malicious, and there are 100 decision makers (DMs). Each of the recommenders is informed about the true state $\theta \in T$, and recommends an action $s_i \in T$. The DMs are informed of which state is the most recommended (i.e., recommended by the majority of recommenders), and each one*

selects an action $\lambda_i \in T$. The payoff of each DM and each of the honest recommenders equals the number of DMs who choose θ. The payoff of the malicious recommender equals the number of DMs who fail to choose θ.

We consider the straightforward equilibrium, SF, in which each of the honest recommenders recommends the true state θ, the malicious recommender recommends the false state, and each of the DMs chooses the most recommended action. It is easy to see that $D(SF) = 1$, because a defection by one honest recommender may push the DMs to switch to the minority recommendation.

As the reader can see, if we alter the game to have four honest recommenders, $D(SF)$ would be 2, because now it would take a defection by two honest recommenders to convince the DMs to switch to the minority recommendation.

The fact that the SF equilibrium in the altered game is more sustainable than in the unaltered game is somewhat surprising, since the two games are "common-knowledge equivalent." It reveals a concern not captured by standard Bayesian equilibria, namely, that one of the two recommenders in the unaltered game may fail to play the equilibrium for the various reasons discussed in the introduction. This concern is reduced in the altered game, in which two recommenders have to fail. Notice also that the malicious recommender can decrease the sustainability of the SF equilibrium by increasing the number of malicious recommenders.

It seems that the sustainability of the equilibrium may itself be subject to strategic considerations. This is a partial explanation of why political parties send many representatives to repeat the same exact "talking points" in different public presentations.

7. Simple rational reasoning supports all index values

The next example shows that any integer, $0, 1, ..., n$, is a possible value of a defection index of some equilibrium of an n-person game. Moreover, these levels can be arrived at by simple rational reasoning, compatible with the computational ability of most players.

Example 7. ***Ride-Sharing Game****: Eight individuals from a small town sign up to attend an event at a specific time and place in a nearby city. For transportation, they each have to sign up and commit to one of two options: riding a private taxi that costs $\$80$, or sharing a ride on a bus that can comfortably take any number of them. The cost of the bus, $\$180$, will be shared equally by the riders who sign up for it. Assuming that the only consideration of every rider is to minimize her transportation cost, what would they choose?*

The table titled "Defection-deterrence computations of eT" illustrates the simplicity of computing the defection index for the profile in which everybody chooses the taxi, eT, as a function of the cost of riding the taxi, c.

When $c = 80$, as in the example above, we look at the third row of the table, the case in which $60 < c \leq 90$. A single player's defection from a taxi to the bus can result only in a loss, raising her cost from \$80 to \$180, so eT is a Nash equilibrium. But what about multiplayer defections?

Even in the presence of one other potential defector, a player who switches from the taxi to the bus is sure to lose: her best possible outcome after switching is a cost of \$90 (= \$180/2) for the bus, instead of \$80 for the taxi. But eT does not deter defections by groups of three or more players, because now a defection may actually improve the player's outcome, by reducing her costs from \$80 for the taxi to \$60 (= \$180/3) for the bus. This leads us to say that eT deters two defectors but does not deter three defectors, and to conclude

Defection-deterrence computations of eT

# of bus riders x	cost/rider $180/x	cost of taxi $c	Deterrence D(eT))
		$180 < c$	0
1	180	$90 < c \leq 180$	1
2	**90**	$60 < c \leq 90$	2
3	60	$45 < c \leq 60$	3
4	45	$36 < c \leq 45$	4
5	36	$30 < c \leq 36$	5
6	30	$25.7 < c \leq 30$	6
7	25.7	$22.5 < c \leq 25.7$	7
8	22.5	$c \leq 22.5$	8

that the deterrence index of eT is $D(eT) = 2$. The table shows how this reasoning applies to all values of taxi costs to generate all the levels of defection deterrence, $0, 1, ...8$.

The eT equilibrium also illustrates the type of defections that players may consider. Do they consider individual or coordinated defections? Three people moving from the taxi line to the bus may come about through explicit communication: for example, when a ride seeker standing in a taxi line approaches two others to coordinate a money-saving joint defection. But alternatively, a ride seeker may switch to the bus, counting on the likelihood that, for similar reasons, at least two others will switch. In either case, having a higher D value, i.e., requiring the participation of more switchers, makes such defections less likely.

8. Forming, switching, and undoing equilibria

8.1. **Entrepreneurial uses of the formation index.** In the ride-sharing game above, it was clear that defecting from the taxi to the bus would be profitable to three or more riders.

Alternatively, direct consideration of the equilibrium in which everybody takes the bus, eB, reveals its low formation value, $F(eB) = 2$. This is useful information for the bus company, which can guide an equilibrium formation

process to eB through the use of sales, promotions, etc. For example, if the bus company guarantees the first two bus-riding candidates that their cost will never exceed \$79, no matter what the other riders do, they can count on the rationality of the first two to take the bus and on the others to follow. Notice that this manipulation by the bus company involves little risk, due to the low formation index, $F(eB) = 2$.

A similar use of the formation index was mentioned in our Example 4, in which players choose whether to subscribe to the new communication network. Recall that the potential pool of subscribers was 200M players, the individual cost of subscription was \$9.99, and the benefit of a subscriber was $k - 9.99$, where k is the number of opponents who subscribe (a zero payoff for players who do not subscribe).

The network provider may be interested in the equilibrium in which every player subscribes, $eSub$. It is easy to see that $F(eSub) = 10$, i.e., if only ten people subscribe, subscribing becomes a dominant strategy for the rest. This information suggests some relatively easy ways to launch the use of the network.

Both examples above illustrate the importance of the formation index for entrepreneurs, policy makers, regulators, and others. Moreover, the minimal information needed for the computations above suggests that it may be easy to compute F, or just to find useful bounds for it in more complex games. This is especially important now, due to the explosion of web devices that gives rise to many new possible games and equilibria.

8.2. **Nonviable switch: changing the US measurement system.** Consider a large game of matching measurement systems, similar to the (language) matching game but in which each player has to select one of the two choices: the metric system, MT, or the US measurement system, US. The two Nash equilibria, namely, eMT, in which every player chooses MT, and eUS, in

which every player chooses US, are both difficult to form, as indicated by their high formation index, $F = 100M$. Thus, establishing either one of these equilibria in a new population would be challenging.

Moreover, in an existing population in which the equilibrium eUS is already established, forming (switching to) the equilibrium eMT would be even more challenging. For every player i and every profile of opponents' choices, β_{-i}, the gain from choosing MT over US, $u_i(MT_i, \beta_{-i}) - u_i(US_i, \beta_{-i})$, is lower in the established population due to the associated transition costs and other such considerations. This means that in this population, defections from MT to US are easier to make than defections from US to MT, which means that $F(eMT) > 100M$. In other words, it is even harder to move a population in which everybody uses US to MT than it is to guide a *new* population to have everybody choose MT.

The US experience with measurement systems illustrates this type of difficulty. Attempts to switch the US population from the use of the US measurement system to the metric system keep failing despite strong encouraging actions taken by the US government and the Congress, in 1866, 1873, 1893, 1968, 1975, and 1988. It seems that encouraging a change does not overcome the high F value. What would be helpful – and perhaps indispensable – is a law imposing penalties for use of the US system. With sufficiently high penalties, the use of the metric system would become a dominant strategy with minimal formation difficulty.

9. Implementation in the presence of faulty players

Eliaz (2002) studies implementation in an environment in which k of n players may be faulty. Implementation in such an environment is difficult because (1) as faulty players, they are irrational and choose unpredictable strategies, and (2) the identity of the k faulty players is unknown. Thus, an

"Eliaz implementor" can rely on the rational behavior of only $n-k$ unknown players. And like the implementor, every rational player knows that she is making choices in an environment with k unknown faulty players.

Eliaz (2002) succeeds in providing a Maskin (1999) type of implementation method for this highly challenging environment by making use of Nash equilibria that he calls *k-fault-tolerant Nash equilibria* (k-$FTNE$). A profile π is a k-$FTNE$ if playing π_i is a best response for every rational player i when it is common knowledge that the number of faulty players is at most k. Through the use of this concept, Eliaz (2002) shows that the implementor can accomplish his goal, provided that the social-welfare function satisfies appropriate monotonicity conditions.

While the objectives of the current paper are different from the objective of Eliaz (2002), the viability indices discussed here provide a simple interpretation of Eliaz's findings. In particular, Eliaz's faulty players may be viewed as a specific type of defectors in the current paper, and his equilibrium concept may therefore be stated through the notion of Nash critical mass presented here. More specifically, saying that "π is k-$FTNE$" in Eliaz' language is the same as saying in the terminology of this paper that "the number of rational players exceeds the Nash critical mass of π," i.e., $n-k \geq NCM(\pi) = F(\pi)+1$. This means that n- $F(\pi) \geq k+1$, i.e., that $D(\pi) \geq k+1$. So π is k-$FTNE$ if and only if $D(\pi) > k$.

We conclude that in an environment with faulty-players, Eliaz implementors can implement a social-welfare function that satisfies Eliaz's monotonicity condition, provided that they use an equilibrium π with a $D(\pi)$ that strictly exceeds the number of faulty players.

As already noted, the current paper deals with issues that cannot be address by the model in Eliaz (2002). While the faulty players of Eliaz are a special

type of defectors in the current paper, the current paper also studies the incentives of rational (nonfaulty) players to defect or to join an equilibrium. As examples in this paper have shown, we can address questions such as when rational players choose to defect, how to modify a game to turn potential defectors into loyalists and vice versa, and how to form or switch an equilibrium in a game played by rational players.

The relationship of this paper to the implementation problem studied by Abraham et al. (2006) is similar to its relationship to Eliaz (2002). Abraham et al. (2006) study implementation in secret-sharing and in multiplayer computation games, and use what they call *resilient equilibria* in order to overcome difficulties due to faulty players. As stated in the introduction, the index of defection deterrence, D, in this paper is the same as the level of *resilience* they use; and in the language of this paper their results show the existence of classes of games that have equilibria with high D values.

Other positive results related to faulty play are presented in Gradwohl and Reingold (2014), who use results about the robustness of equilibria of large games (see Kalai (2004)) to show that such equilibria can sustain a significant number of defectors.

10. Matching in networks

Games on networks, as in Jackson and Zenou (2015), provide an understanding of the viability of Nash equilibria as determined by social connectivity. It is easy to compute the defection index for equilibria of a *network matching game*, described as follows.

The set of players N consists of the vertices in a graph with a set of directed edges $E \subset \{(i,j) \in N \times N : i \neq j\}$. The set of (outward-directed) *neighbors* of a player i is defined by $\eta b(i) = \{j \in N : (i,j) \in E\}$. Every player selects

a choice X from a set of possible choices, and her payoff is the number of her neighbors that her choice matches.

Let C denote the set of *connected* players, i.e., players i with $\eta b(i) \neq \emptyset$, and think of any player $v \in C$ as *most vulnerable*, if she is minimally connected among all connected players, i.e., $|\eta b(v)| = \min_{j \in C} |\eta b(j)|$.

Proposition 5. *The **defection index in network matching games**: The defection index of the profile in which every player chooses X, eX, is:*

$D(eE) = n$ if no player is connected, i.e., $C = \varnothing$; otherwise,

$D(eE) = [|\eta b(v)|/2] + 1$, where v is any most vulnerable player. In other words, $D(eE)$ is the strict majority of the neighbors of a least connected player.[13]

Proof. Let M denote the strict majority of the neighbors of a most vulnerable player. If the number of eE defectors at a profile α is strictly smaller than M, then eE must be a best response to α. So $D(eE) \geq M$.

Conversely, consider a profile β in which all but M of the neighbors of some most vulnerable player choose E and the rest of the neighbors choose the same alternative A; then eE is not a best response to β. It follows that $D(eE) = M$. □

A similar analysis can easily be conducted for problems of mismatching in a network. For illustration, consider the divisive equilibrium Div of the Party Line game of the introduction, in which all three Democrats choose F and all five Republicans choose E. This game may be described by a bipartite graph, connecting every player to all the players of the opposite party, with a player's payoff being the number of her neighbors that her choice *mismatches*. Following the logic above, every Republican is most vulnerable, since he is connected only to the three Democrats, whereas each Democrat is connected

[13]Recall that [x] is the largest integer that is strictly smaller than x.

to the five Republicans. Conducting the same analysis as above, we conclude that $D(Div) = [3/2] + 1 = 2$.

10.1. **Centralized vs. decentralized interaction.** The Language Choice Game, Example 2, is a network matching game based on a complete graph, i.e., every two distinct players are connected (in both directions). In this game every player has $200M - 1$ neighbors, so $D(eE) = [(200M - 1)/2] + 1 = 100M$.

This is in contrast to the low defection-deterrence level of centralized interaction, as discussed in the Match-the-Center game, Example 5. That game is based on a star-shaped graph in which the boss, B, is the center vertex, and n subordinate players are vertices connected only to him. In that game the subordinates are the most vulnerable players, as they are each connected only to one player, and the equilibrium eE has the minimal value $D(eE) = 1$.

The contrast in defection-deterrence levels of the two games just cited has implications in a variety of contexts. In a political context, it suggests that matching equilibria are significantly more sustainable (in the sense defined in this paper) in free societies than in dictatorships. For games of currency choices, it suggests that a free-trade equilibrium is more sustainable than a centralized-trade equilibrium. In supply-chain games it means that relying on a single source is risky, and backup sources for supplies are important for sustainability.

As already discussed, the low defection-deterrence value of the star-shaped graph is due to the total dependence of the subordinate players on the boss, B. If B defects from the equilibrium – due, for example, to threats and bribes or to miscalculations – the effect on all the subordinates may be devastating.

11. Future research

11.1. **Expanding Nash's theorem.** Nash's existence theorem provides sufficient conditions for the existence of a Nash equilibrium, i.e., a profile π with

$D(\pi) \geq 1$. For applications in which the low level of sustainability $D(\pi) = 1$ is unsatisfactory, it is important have sufficient conditions for the existence of a profile π with $D(\pi) \geq k$, for $k = 2, 3, ..., n$.

11.2. **More refined indices.** As discussed in the introduction, in order to stay within the Nash equilibrium approach, the viability indices discussed in this paper deal only with ordinal anonymous defections. i.e., they consider only whether players may gain a positive payoff by defection, disregarding the amount that they may gain; and they only study the number of potential defectors, disregarding the identity and role of the defectors in the game.[14]

The severe limitation of the ordinality assumption can be seen in the Confession game, Example 3. If the seven-year sentence reduction awarded to a confessor were changed to a one-hour sentence reduction, the unviable no-confession equilibrium (with $D(nC) = 1$) should become more viable and be assigned a significantly higher D value, unlike our ordinal index.

The importance of the role of defectors in the game is illustrated next by considering the common language equilibrium, eE, of Example 2, played in the following three communication graphs:

(1) K_n is the complete graph on n players;

(2) K_n^+ amends K_n by the addition of one player H who is connected only to one of the original players W; and

(3) S_n is the star-shaped graph in which n players are each connected to only one central player.

The significant discrepancy between $D_{K_n}(eE) \approx n/2$ and $D_{K_n^+}(eE) = 1$ ignores the fact that to all but one player, the reliability of communications in the two networks are the same. This discrepancy suggests the construction of viability indices that average viability levels from individual points of view.

[14]Substantial discussion of this subject with highly revealing examples is provided by Goeree and Holt (2001) and their follow-up papers.

A disturbing lack of discrepancy can be seen when we compare $D_{K_n^+}(eE) = 1$ with $D_{S_n}(eE) = 1$. In both games one player's defection can undo the equilibrium. But the defection of the central player in S has devastating consequences to the rest of the players, whereas the defection of W in K_n^+ is significant to only one player.

The structural issues illustrated above suggest a need for nonanonymous indices that take into account defecting players' position in the game under consideration.

11.3. **Nonanonymous indices and equilibrium formation.** Nonanonymous indices may be more applicable, but may also require more demanding computations. We proceed to show how concepts discussed in this paper may fit into such a broader discussion.

Our anonymous formation index $F(\pi)$ is the minimal number of π loyalists that guarantees the formation of π. But as discussed below, if we can target and select the players who form π, its formation may require a smaller number.

Recall our earlier definition that a coalition C *forms* the equilibrium π if the play of π by C makes π a dominant strategy for the remaining players; and that π-formation defines a monotonic partial order over the coalitions of the game (if C forms π, so does any of its supersets), with the grand coalition N being its unique maximal element.

Since our underlying game has a finite number of players, we can identify the minimal forming coalitions – the *roots*.

Definition 7. *A* π-root *is a minimal* π*-forming coalition, i.e., no strict subset of the coalition forms* π*.*

Clearly, a coalition forms π if and only if it contains a root, and a coalition is incentivized to play π if and only if its complement contains a root. Also, any coalition C of size $|C| \geq F(\pi)$ must contain a root.

Consider, for example, the divisive equilibrium, Div, of the Party Line game in which the three Democrats all choose F and the five Republicans all choose E, Example 1. It is easy to verify that Div has the following roots: (i) the coalition of all the Democrats, Ds, (ii) the coalition of all the Republicans, Rs, and (iii) any coalition that consists of three Republicans and two Democrats, $C_{3Rs,2Ds}$.

With this in mind, we may consider the following four ways of forming the equilibrium Div:

CDs: Convince the three Democrats to choose F.

CRs: Convince the five Republican to choose E.

C$C_{3Rs,2Ds}$: Convince any group of 3 Republicans and 2 Democrats to choose their divisive strategies.

C6: Convince any six players to choose their divisive strategies.

CDs, CRs, and C$C_{3Rs,2Ds}$ work, because they target roots by name. C6 is anonymous and it works because $6 = F(Div)$.

As this example illustrates, it may be more efficient to form an equilibrium through the use of nonanonymous coalitions. But due to the multiplicity of roots, the decision of which root to target may require more information and computations.

11.4. **Sequential formation processes.** Sequential formation processes may be developed by replacing the one-step formation with multistage ones.[15] For example, the Div equilibrium of the Party Line game may be triggered by recruiting only two of the Democrats to choose F; this in turn will incentivize the five Republicans to choose E, which in turn will incentivize the remaining (third) Democrat to choose F.

[15]For a more general study of dynamic processes of equilibrium formation, we refer the reader to Chwe (1994).

While the more refined multistage approach is useful in many applications, it involves a larger number of possible procedures, which in turn requires more complex computations. Indeed, the genius of dictators lies in their ability to navigate a multistage formation process leading to an equilibrium in which all the players follow their wishes.

Computations of multistage equilibrium formation are related to computations of sequential elimination of dominated strategies (see, for example, Gilboa et al. (1993) and Marx and Swinkels (1997)). Indeed, as shown in Gilboa et al. (1993), such computations are NP-complete.

12. References

Abraham, I., D. Dolev, R. Gonen, and J. Halpern (2006), "Distributed computing meets game theory: robust mechanisms for rational secret sharing and multiparty computation," in Proceedings of the. 25th ACM Symposium on Principles of Distributed Computing, 53–62.

Aumann, R.J. (1959), "Acceptable points in general cooperative n-person games," *Contributions to the Theory of Games IV*, Annals of Mathematics Study 40, Princeton University Press, 287-324.

Aumann, R.J. and S. Sorin (1989), "Cooperation and bounded recall," *Games and Economic Behavior*, 1 (1), 5-39.

Basu, K. and J.W. Weibull (1991), "Strategy subsets closed under rational behavior," *Economics Letters* 36, 141-146.

Ben-Or, M., S. Goldwasser, and A. Wigderson, (1988) "Completeness theorems for non-cryptographic fault-tolerant distributed computation," in STOC '88 Proceedings of the twentieth annual ACM symposium on Theory of computing, ACM New York, 1-10.

Bernheim, B. D. , B. Peleg, and M. D. Whinston (1987), "Coalition-proof equilibria: I. Concepts," *Journal of Economic Theory*, 42, 1–12.

Camerer, C. (2003), *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton, NJ: Princeton University Press.

Chwe, M. (1994), "Farsighted coalitional stability," *Journal of Economic Theory*, 63, 299-325.

Crawford, V.P. (1998), "A Survey of Experiments on Communication via Cheap Talk." *Journal of Economic Theory*, 78, 286-298.

Crawford, V.P. and J. Sobel (1982), "Strategic information transmission," *Econometrica,* 50 (6), 1431-1451.

Crawford, V.P., Costa-Gomes, M.A., and N. Iriberri (2013), "Structural models of nonequilibrium strategic thinking: theory, evidence, and applications," *Journal of Economic Literature,* 51 (1), 5–62.

Eliaz, K. (2002), "Fault-tolerant implementation," *Review of Economic Studies*, 69(3), 589-610.

Ellison, G. (2000), "Basins of attraction, long run stochastic stability, and the speed of step-by-step evolution," *Review of Economic Studies*, 67 (1), 17-45.

Erev, I. and A. E. Roth (1998), "Predicting how people play games: reinforcement learning in experimental games with unique,mixed strategy equilibria," *The American Economic Review,* 88 (4), 848-881.

Gilboa, I., E. Kalai, and E. Zemel (1993), "The complexity of eliminating dominated strategies," *Mathematics of Operations Research,* 18, 553-565.

Goeree, J.K. and C.A. Holt (2001), "Ten little treasures of game theory and ten Intuitive contradictions," *American Economic Review*, 91 (5), 1402-1422.

Gradwohl, R. and O. Reingold (2014), "Fault tolerance in large games," *Games and Economic Behavior*, 86, 438-457.

Jackson, M. O. and Y. Zenou. (2015) "Games on Networks," In *Handbook of Game Theory with Economic Applications.* Vol. 4. Elsevier.

Kahneman, D. and A. Tversky (2000), *Choices, Values, and Frames*, New York : Russell Sage Foundation.

Kalai, E. and A. Neme (1992), "The strength of a little perfection," *International Journal of Game Theory*, 20 (4), 335-355.

Kalai, E. and D. Samet (1984), "Persistent equilibria in strategic games," *International Journal of Game Theory*, 13(3), 129-144.

Kalai, E. (2004), "Large robust games," Econometrica, 72 (6), 1631-1665.

Kandori, M., G. Mailath, and R. Rob (1993) "Learning, mutation, and long run equilibria in games," *Econometrica*, 61(1), 29-56.

Kohlberg, E. and J.F. Mertens (1986), "On the strategic stability of equilibria," *Econometrica*, 54 (5), 1003-1037.

Kreps, D.M. and R. Wilson (1982), "Sequential equilibria," *Econometrica,* 50 (4), 863-894.

Marx, L. M., and J. M. Swinkels (1997), "Order independence for iterated weak Dominance," *Games and Economic Behavior*, 18, 219-245.

Maskin, E. (1999), "Nash implementation and welfare optimality," *Review of Economic Studies*, 66, 23-38.

Mauersberger, F. and R. Nagel (2018), "Levels of Reasoning in Keynesian Beauty Contests: A Generative Framework," Handbook of Computational Economics, Volume 4, Heterogeneous Agents, Cars Hommes and Blake LeBaron editors, Amsterdam: North-Holland.

Mertens, J.F. (1992), "The small worlds axiom for stable equilibria," *Games and Economic Behavior*, 4, 553-564.

Mishina, K. (1992), "*Toyota Motor Manufacturing, U.S.A., Inc.*" Harvard Business School Case 693-019, September 1992. (Revised September 1995.)

Moreno, D. and J. Wooders (1996), "Coalition-proof equilibrium," *Games and Economic Behavior*, 17, 80–112.

Myerson, R.B. (1978), "Refinements of the Nash equilibrium concept," *International Journal of Game Theory,* 7, 73-80.

Myerson R.B. and J.W. Weibull (2015), "Tenable strategy blocks and settled equilibria," *Econometrica* 83 (3), 943-976.

Nagel, R. (1995), "Unraveling in guessing games: an experimental study," *American Economic Review*, 85 (5), 1313-1326.

O'Neill, B. (1987), "Nonmetric test of the minmax theory of two-person zerosum Game," *Proceedings of the National Academy of Sciences*," 84(7), 2106– 09.

Schelling, T.C. (1960), *The strategy of conflict* (1st ed.), Cambridge: Harvard University Press.

Selten, R. (1975), "Reexamination of the perfectness concept for equilibrium points in extensive games," *International Journal of Game Theory*, 4 (1), 25-55.

Smith V.L. (1982), "Microeconomic systems as an experimental Science," *The American Economic Review*, 72 (5), 923-955.

Young, P. (1993), "The evolution of conventions," *Econometrica*, 61 , 57-84.

Kellogg School of Management, Northwestern University

E-mail address: kalai@kellogg.northwestern.edu

URL: http://www.kellogg.northwestern.edu/faculty/directory/kalai_ehud.aspx